A SAS/IML Program to Implement a General Approximate Degrees of Freedom Solution for Inference and Estimation

Lix and Keselman (1995) presented a program written in SAS/IML (SAS Institute Inc., 2004) to test omnibus and subeffects general linear model hypotheses using a general approximate degrees of freedom (ADF) statistic ($TWJ/c$). Keselman, Wilcox and Lix (2003) modified this program to obtain numeric results for the general ADF solution using robust estimators ($TWJt/c$), that is, trimmed means and Winsorized variances (covariances) and a bootstrap method to obtain an empirical critical value to evaluate the test statistic.

A new version of this program is described in the following documentation. Like previous versions, the new program conducts tests of omnibus (i.e., overall) main and interaction effects in addition to tests of individual linear contrasts or families of contrasts for independent groups and correlated groups designs. The primary addition to this new program is that it computes robust effect size (ES) estimates and robust confidence intervals (CIs) for these estimates.

This document has the following components:

- A description of the program inputs and default options, and a listing of the program code.
- A numeric example for a one-way independent groups design.
- A numeric example for a two-way independent groups design.
- A numeric example for a correlated groups design.
Program Inputs and Default Options

The main program module is called **WJGLM**. It is invoked with a **RUN WJGLM** statement. It requires the following inputs: **Y**, **C**, **NX**, **OPT1**, **OPT2**, and **OPT3**. Depending on the values for **OPT1**, **OPT2**, and **OPT3**, other, optional inputs may be specified. The program output is determined by the inputs. A brief description of the required and optional inputs follows.

- **Y** is the input data matrix. (REQUIRED)
- **NX** is a vector of the number of observations in each group or cell of the design. (REQUIRED)
- **C** is a matrix that defines linear contrasts among the levels of the between-subjects factor(s). The dimensions of **C** depend on the hypothesis being tested and the number of between-subjects factors in the design. (REQUIRED)
- **U** is a matrix that defines linear contrasts among the levels of the within-subjects factor(s). The dimensions of **U** depend on the hypothesis being tested and the number of within-subjects factors in the design. (REQUIRED FOR DESIGNS CONTAINING WITHIN-SUBJECTS FACTORS)
- **OPT1** is a scalar indicator to control the use of robust estimators. It takes a value of 0 or 1. A 0 indicates the program user wants to compute the ADF test statistic with the usual least-squares estimators, while a 1 indicates that trimmed means and Winsorized variances/covariances will be used to compute the ADF test statistic. (REQUIRED).
- When $\text{OPT1} = 1$, the program user can specify $\text{PER}$. (OPTIONAL) $\text{PER}$ is the proportion of trimming in each tail of the data distribution. The minimum value of $\text{PER}$ is 0 (no trimming) and the maximum plausible value is 0.49 (i.e., 49% trimming in each tail), although we recommend the default $\text{PER} = .20$, which represents 20% symmetric trimming (i.e., 20% of the observations in each tail are trimmed).

- $\text{OPT2}$ is a scalar indicator to control the use of the bootstrap to compute a critical value for the test statistic produced by $\text{WJGLM}$. It takes a value of 0 or 1. A 0 indicates that the program user does not want to use the percentile bootstrap method to compute a critical value for the test statistic, $T_{\text{WJ}}/c$ (or $T_{\text{WJt}}/c$), but instead wants to adopt a theoretical critical value, while a 1 indicates that the non-parametric percentile bootstrap method is used to obtain an empirical critical value. (REQUIRED)

- When $\text{OPT2} = 1$, the program user can also specify values for $\text{NUMSIM}_B$ and $\text{SEED}$. (OPTIONAL) $\text{NUMSIM}_B$ is a positive integer that defines the number of bootstrap samples to generate a critical value when $\text{OPT2} = 1$. The program user can specify any value for $\text{NUMSIM}_B$, but the default is 999. Carpenter and Bithell (2000) recommend between 1000 and 2000 samples for setting CIs around a parameter estimate, but Efron and Tibshirani (1986) note that for other applications of the bootstrap, such as computing the standard error, as few as 100 bootstrap samples will be sufficient. Choosing a value of $B$ such that $(B + 1)\alpha$ is an integer value is recommended, because it
avoids the need for interpolation. **SEED** is a positive integer less than $2^{31} - 1$ that defines the initial value for the call to the random number generator for the bootstrap simulations. The default is **SEED = 0**, which specifies a value from the computer's internal clock.

- **OPT3** is a scalar indicator to control the production of ES estimates and CIs. It takes a value of 0 or 1. A 0 indicates that the program user does not want to compute an ES estimate. A 1 indicates that an ES estimate and its CI should be computed using the percentile bootstrap method. (REQUIRED)
  - When OPT3 = 1, the program user can specify values for **NUMSIM_ES**, **SEED**, **ALPHA**, **LOC1**, **LOC2**, and **SCALE** (OPTIONAL). **NUMSIM_ES** is a positive integer that defines the number of bootstrap samples to generate a CI for the ES estimate. The default is **NUMSIM_ES = 999**. **SEED** is a positive integer less than $2^{31} - 1$ that defines the initial value for the call to the random number generator for the bootstrap simulations. The default is **SEED = 0**, which specifies that a value from the computer's internal clock is used as the initial value. **ALPHA** defines $1 - \alpha$, the probability coverage when computing the CI. The default is **ALPHA = .05**. **LOC1** and **LOC2** are “locator” indicators. These integers are used to select the standardizer for the ES estimate and can take on any value assigned by the user. For factorial or multivariate designs, **LOC1** will select the row in which the standardizer is located, while **LOC2** will select the column in which the standardizer is located. The default values are **LOC1 = 1** and **LOC2 = 1**. 
LOC2 = 1, which indicate that the first group (or cell) of the design will be selected as the standardizer. If the user specifies LOC1 = 0 and LOC2 = 0, the square root of the average of the variances over the cells involved in the contrast is used as the standardizer. If the user specifies LOC1 = 99 and LOC2 = 99, no standardizer is selected.

SCALE is a scalar indicator to control the use of a scaling factor for the effect size estimator (i.e., .642 for 20% symmetric trimming) when robust estimators are adopted. It takes a value of 0 or 1; a zero indicates that no scaling factor will be used, while a 1 indicates that a scaling factor will be adopted. The default is SCALE=1.

- Additional program details:
  - A single ADF test statistic or ES estimate is produced with each call to WJGLM.
  - Another module, called BOOTCOM, is invoked to compute the ADF test statistics for a family of contrasts/comparisons when the bootstrap is used to obtain a critical value to control the familywise error rate (FWER). This module is initiated with a RUN BOOTCOM statement. One optional input for BOOTCOM is ALPHA, which sets the FWER. The default value is ALPHA = .05. A second optional input for BOOTCOM is NUMSIM_BC, the number of bootstrap samples for computing a critical value for FWER. The default is 699; this value was selected based on the guidelines of Efron and Tibshirani (1986) and our experience in implementing the program, which is more
computationally complex than the WJGLM module and can therefore result in errors due to memory size on some computers. In order for **BOOTCOM** to be invoked, C and U must specify a set of contrasts among the levels of the between-subjects and within-subjects factors(s).

- Note that this program uses the same starting seed for the bootstrap for both **OPT2** (i.e., for inference about a test statistic) and **OPT3** (i.e., for computing empirical CIs for ESs). **SEED** only needs to be specified once after you run the SAS/ program code; new seeds will be automatically generated for each subsequent call to WJGLM or **BOOTCOM**. The advantage of specifying a starting seed is that the results produced by this SAS program can be replicated. If replication of results is not a critical issue, then the default value can be used.

- Note that the program will not compute an ES estimate using the noncentral $t$ distribution, for the reasons discussed in the accompanying paper.
References


